Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет Информационных технологий
Кафедра Информационных систем и технологий
Специальность 1-98 01 03 «Программное обеспечение информационной безопасности мобильных систем»

ОТЧЁТ
ПО ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине «Теория вероятности математическая статистика»
Тема: «Линейная регрессия. Криволинейная регрессия»

Исполнитель:
студент 2 курса 8 группы
Солодкий Д. В.
Руководитель:
Волк А. М.

Минск 2022

Лабораторная работа 2.
Линейная регрессия. Криволинейная регрессия

Пусть изучается связь между двумя величинами на основании экспериментальных данных – по выборке объема n пар значений: $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$.

Две случайные величины (СВ) могут быть: 1) независимыми; 2) связаны функциональной зависимостью (каждому значению одной из них соответствует строго определенное значение другой); 3) связаны статистической зависимостью (каждому значению одной СВ соответствует множество возможных значений другой и изменение значения одной величины влечет изменение *распределения* другой).

При изучении статистической зависимости обычно ограничиваются исследованием усредненной зависимости: как в среднем будет изменяться значение одной величины при изменении другой. Такая зависимость называется ***регрессионной***.

Основным методом исследования статистических зависимостей является ***корреляционно-регрессионный*** анализ.

Основными задачами корреляционного анализа являются выявление связи между наблюдаемыми СВ и оценка тесноты этой связи.

Основными задачами регрессионного анализа являются установление *формы зависимости* между наблюдаемыми величинами и определение по экспериментальным данным уравнения зависимости, которое называют ***выборочным (эмпирическим) уравнением регрессии***, а также прогнозирование с помощью уравнения регрессии среднего значения зависимой переменной при заданном значении независимой переменной.

Вид эмпирической функции регрессии определяют исходя из: 1) соображений о физической сущности исследуемой зависимости; 2) опыта предыдущих исследований; 3) характера расположения точек на ***корреляционном поле***, которое получается, если отметить на плоскости все точки с координатами (x_i, y_i), соответствующие наблюдениям.

Наибольший интерес представляет линейное эмпирическое уравнение регрессии $\hat{y} = b_0 + b_1 x$, так как: 1) это наиболее простой случай для расчетов и анализа; 2) при нормальном распределении функция регрессии является линейной.

Количественной мерой *линейной связи* между двумя наблюдаемыми величинами служит ***выборочный коэффициент корреляции***.

$$r_{x;\,y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{D_{\text{в}}(x) D_{\text{в}}(y)}},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ – выборочные средние величин x, y и произведения xy соответственно; $D_{\text{в}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$, $D_{\text{в}}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2$ – выборочные дисперсии величин x и y.

Свойства выборочного коэффициента корреляции.

1. $-1 \le r_{x;\,y} \le 1$.

2. Если наблюдаемые величины x и y независимы, то $r_{x;\,y} \approx 0$. Однако обратное неверно: значение $r_{x;\,y} \approx 0$ не гарантирует, что наблюдаемые величины x и y независимы.

3. Если $\left|r_{x;\,y}\right| = 1$ (или близок к 1), то наблюдаемые величины x и y связаны линейной зависимостью, т. е. $y = b_0 + b_1 x$.

4. Если $r_{x;\,y} > 0,$ то с ростом значений одной величины значения другой также в основном возрастают; если $r_{x;\,y} < 0,$ то с ростом значений одной величины значения другой, наоборот, убывают.

Проверка значимости коэффициента корреляции – это проверка гипотезы о том, что коэффициент корреляции значимо отличается от нуля. Так как выборка произведена случайно, нельзя утверждать, что если выборочный коэффициент корреляции $r_{x;\,y} \neq 0,$ то и коэффициент корреляции генеральной совокупности $r_{\xi;\,\eta} \neq 0$. Возможно, отличие $r_{x;\,y}$ от 0 вызвано только случайными искажениями наблюдаемых значений.

Если выборка из нормального распределения, то проверка производится по *критерию Стьюдента*: если

$$t_{\text{расч}} = \left|r_{x;\,y}\right| \sqrt{\frac{n-2}{1-r_{x;\,y}^2}} > t_{\text{табл}} = t_{\alpha;\,n-2},$$

где $t_{\alpha;\,n-2}$ – квантиль уровня α распределения Стьюдента с числом степеней свободы $k = n - 2$ (определяется по таблице), то при заданном уровне значимости α (допускается, что вывод может быть ошибочным с небольшой вероятностью α) коэффициент корреляции считается значимо отличающимся от нуля, а следовательно, связь между величинами x, y признается статистически значимой.

Если коэффициент корреляции на основании проверки признается значимо отличающимся от нуля, считают допустимым принять предположение о линейной регрессионной зависимости между наблюдаемыми величинами.

Подчеркнем, что *коэффициент корреляции является мерой именно линейной зависимости*. В случае нелинейной зависимости связь между величиной коэффициента корреляции и близостью точек корреляционного поля к некоторой линии не прослеживается. Поэтому в практических задачах при выборе вида эмпирической функции регрессии обязательно учитывают характер расположения точек на корреляционном поле.

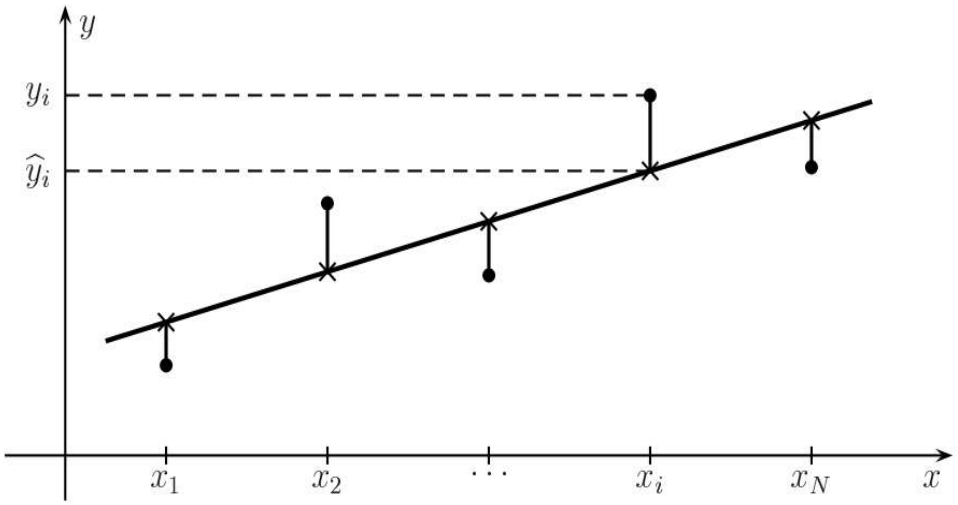
Определение коэффициентов эмпирического линейного уравнения регрессии методом наименьших квадратов. Пусть имеется выборка объема n наблюдений над двумя величинами x и y: $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$, и принята гипотеза о линейной зависимости между y и x. Для определения коэффициентов линейного эмпирического уравнения регрессии

$$\hat{y} = b_0 + b_1 x$$

используется ***метод наименьших квадратов*** (***МНК***). Суть этого метода в том, что коэффициенты b_0 и b_1 выбирают так, чтобы сумма квадратов отклонений

наблюдаемых значений y_i от предсказываемых по уравнению $\hat{y}_i = b_0 + b_1 x_i$ была минимальной (см. рис.). Таким образом, минимизируется функция

$$Q(b_0; b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \to \min_{b_0, b_1}.$$



Согласно МНК, значения параметров b_0 и b_1 находят из системы, которая называется ***системой нормальных уравнений*** метода наименьших квадратов:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Метод наименьших квадратов широко применяется при статистической обработке результатов измерений.

Зависимость между двумя наблюдаемыми величинами далеко не всегда можно выразить линейной функцией. Иногда видно, что точки корреляционного поля образуют некоторую кривую. При выборе вида эмпирической функции регрессии необходимо учитывать теоретические сведения и опыт предыдущих аналогичных исследований.

Как правило, до начала исследования должен быть определен вид эмпирической функции регрессии с точностью до нескольких параметров, значения которых оцениваются по результатам эксперимента. В том случае, если функция регрессии линейна по параметрам или может быть сведена к таковой с помощью замены переменных, для определения оценок параметров используют МНК.

Например, коэффициенты квадратичного уравнения регрессии $\hat{y} = b_0 + b_1 x + b_2 x^2$ находят из следующей системы нормальных уравнений:

$$\begin{cases} nb_0 + b_1\sum\limits_{i=1}^{n} x_i + b_2\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i, \\ b_0\sum\limits_{i=1}^{n} x_i + b_1\sum\limits_{i=1}^{n} x_i^2 + b_2\sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} x_i y_i, \\ b_0\sum\limits_{i=1}^{n} x_i^2 + b_1\sum\limits_{i=1}^{n} x_i^3 + b_2\sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} x_i^2 y_i. \end{cases}$$

Степенная зависимость вида $y = ax^b$ может быть сведена к линейной с помощью логарифмирования:

$$\ln y = \ln a + \ln x^b \quad \Rightarrow \quad \ln y = \ln a + b\ln x.$$

Если ввести новые переменные $Y = \ln y, X = \ln x,$ исходная зависимость сведется к линейной $Y = b_0 + b_1 X,$ коэффициенты которой могут быть найдены по МНК. Тогда коэффициенты искомой зависимости определятся из соотношений $a = \mathrm{e}^{b_0}, b = b_1.$ В таблице приведены некоторые виды зависимостей, которые сводятся к линейной после замены переменных.

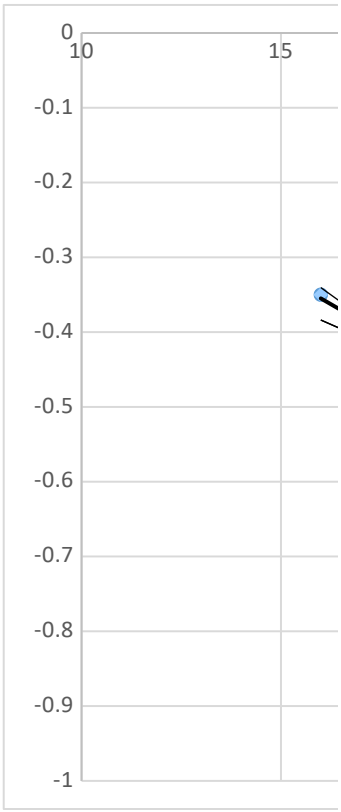
Вид зависимости	Уравнение зависимости	Замена переменных, сводящая зависимость к линейной $Y = b_0 + b_1 X$	Выражение параметров зависимости через коэффициенты b_0, b_1
Гиперболическая	$y = a + \dfrac{b}{x}$	$Y = y, X = \dfrac{1}{x}$	$a = b_0, b = b_1$
Логарифмическая	$y = a + b\ln x$	$Y = y, X = \ln x$	$a = b_0, b = b_1$
Экспоненциальная	$y = a\,\mathrm{e}^{bx}$	$Y = \ln y, X = x$	$a = \mathrm{e}^{b_0}, b = b_1$
Степенная	$y = ax^b$	$Y = \ln y, X = \ln x$	$a = \mathrm{e}^{b_0}, b = b_1$
Гиперболическая	$y = \dfrac{1}{a + bx}$	$Y = \dfrac{1}{y}, X = x$	$a = b_0, b = b_1$

Для проверки того, удачно ли выбран вид зависимости, следует построить новое корреляционное поле на плоскости *OXY*. Если вид зависимости y от x подобран правильно, то точки $(X_i; Y_i)$ будут располагаться вдоль прямой.

Для выбора наилучшей аппроксимирующей функции из нескольких (в случае, когда нет теоретического обоснования для выбора определенного вида зависимости) используют коэффициент детерминации R^2, который принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. В случае линейной зависимости R^2 равен квадрату выборочного коэффициента корреляции.

C	16	23	26	28	30	36	40
T	-0.35	-0.57	-0.61	-0.69	-0.75	-0.81	-0.94

xi	yi	xiyi	xi^2	yi^2
16	-0.35	-5.6	256	0.1225
23	-0.57	-13.11	529	0.3249
26	-0.61	-15.86	676	0.3721
28	-0.69	-19.32	784	0.4761
30	-0.75	-22.5	900	0.5625
36	-0.81	-29.16	1296	0.6561
40	-0.94	-37.6	1600	0.8836
199	-4.72	-143.15	6041	3.3978
n=	7			
r=	-0.98687	tрасч=	13.65984	tтабл= 2.570582



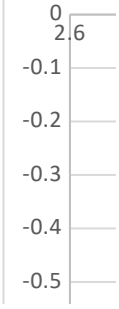
Система нормальных уравнений

7	199		-4.72
199	6041		-143.15

2.249069248	-0.07409	b0=	-0.00993
-0.074087863	0.002606	b1=	-0.02337

X = 1/x	Y=1/y
0.0625	-2.85714
0.043478	-1.75439
0.038462	-1.63934
0.035714	-1.44928
0.033333	-1.33333
0.027778	-1.23457
0.025	-1.06383

X = lnxi	Y	XY	Xi^2	Yi^2
2.7725887222	-0.35	-0.97041	7.687248	0.1225
3.1354942159	-0.57	-1.78723	9.831324	0.3249
3.258096538	-0.61	-1.98744	10.61519	0.3721
3.3322045102	-0.69	-2.29922	11.10359	0.4761
3.4011973817	-0.75	-2.5509	11.56814	0.5625
3.5835189385	-0.81	-2.90265	12.84161	0.6561
3.6888794541	-0.94	-3.46755	13.60783	0.8836
23.171979761	-4.72	-15.9654	77.25494	3.3978
n=	7			
r=	-0.99162	tрасч=	13.72562	tтабл= 2.570582



7	23.17198		-4.72		
23.171979761	77.25494		-15.9654		

20.098051833	-6.02824	b0=	1.380489	a=b0	1.380489
-6.028244642	1.821066	b1=	-0.62072	b=b1	-0.62072

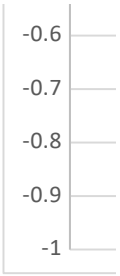
Xi=xi	Xi^2	Xi^3	Xi^4	Yi=yi	XiYi	Xi^2Yi
16	256	4096	65536	-0.35	-5.6	-89.6
23	529	12167	279841	-0.57	-13.11	-301.53
26	676	17576	456976	-0.61	-15.86	-412.36
28	784	21952	614656	-0.69	-19.32	-540.96
30	900	27000	810000	-0.75	-22.5	-675
36	1296	46656	1679616	-0.81	-29.16	-1049.76
40	1600	64000	2560000	-0.94	-37.6	-1504
199	6041	193447	6466625	-4.72	-143.15	-4573.21

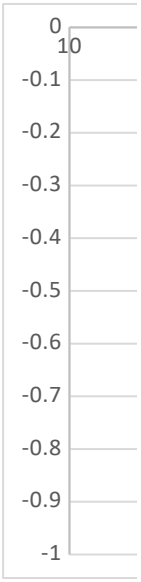
n= 7

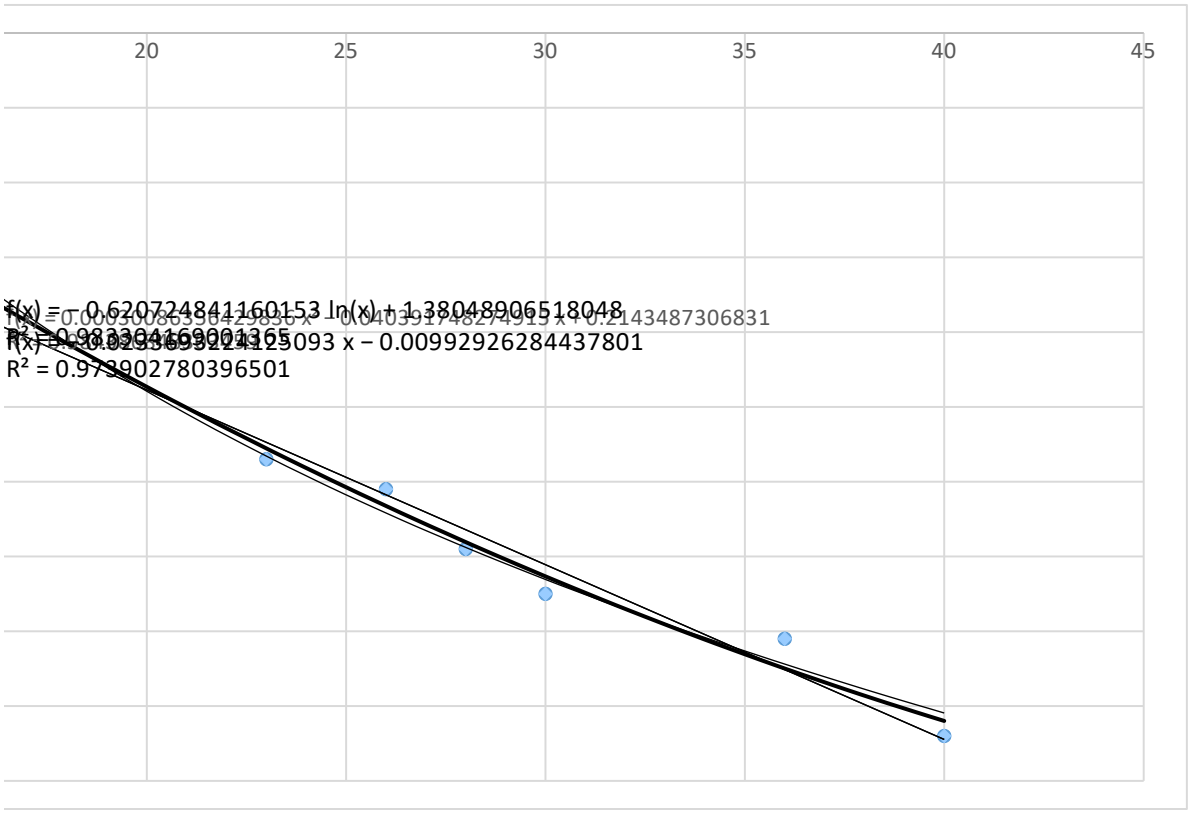
r= -0.98687 tрасч= 13.65984 tтабл= 2.570582

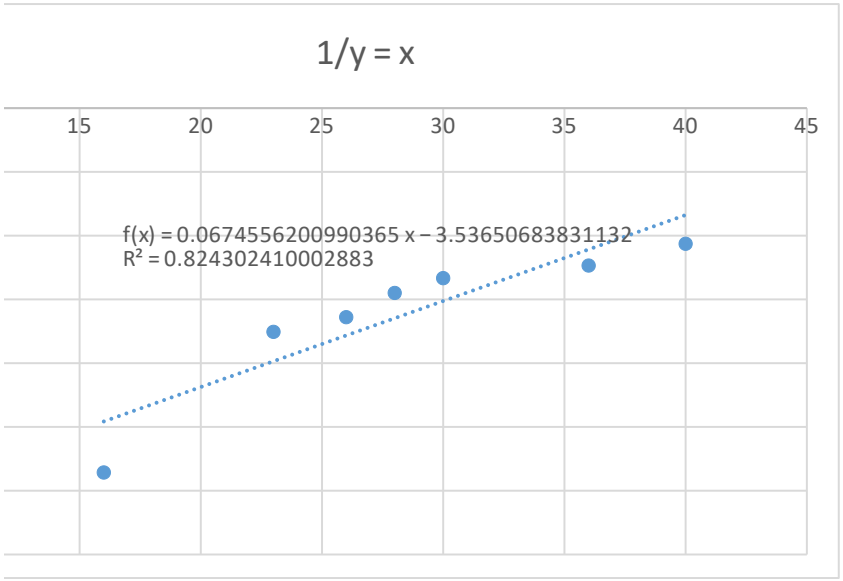
6466625	193447	6041	-4573.21
193447	6041	199	-143.15
6041	199	7	-4.72

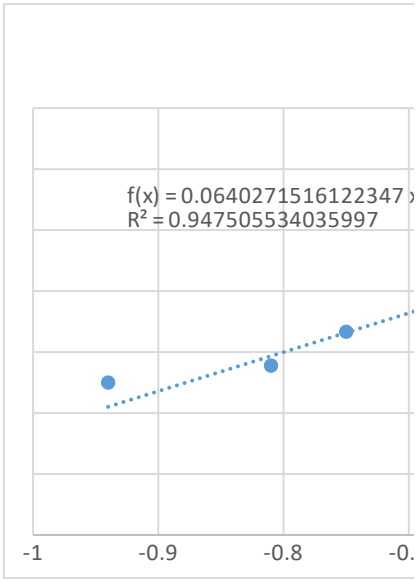
4.01259E-05	-0.00227	0.029912	a=	0.0003008636
-0.002270265	0.131054	-1.76645	b=	-0.0403917483
0.0299117505	-1.76645	24.54671	c=	0.2143487307

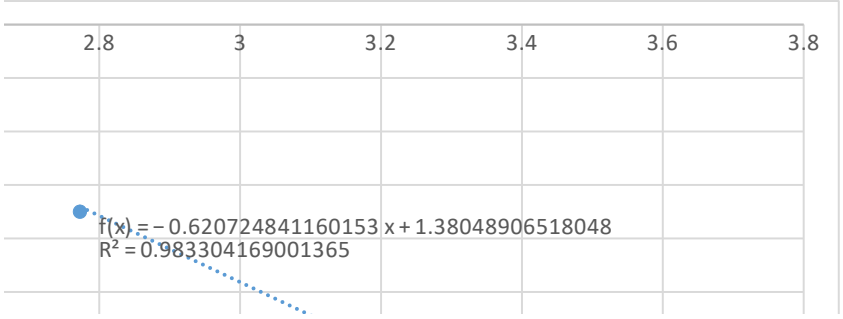


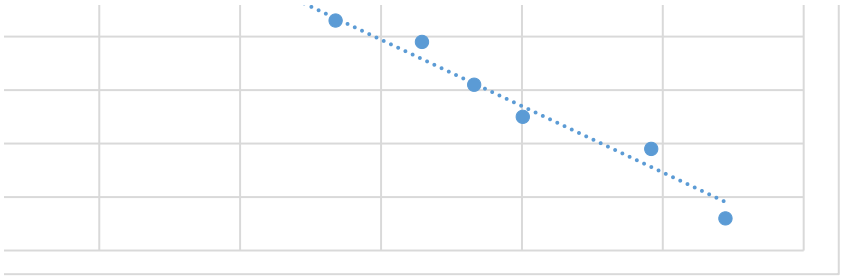


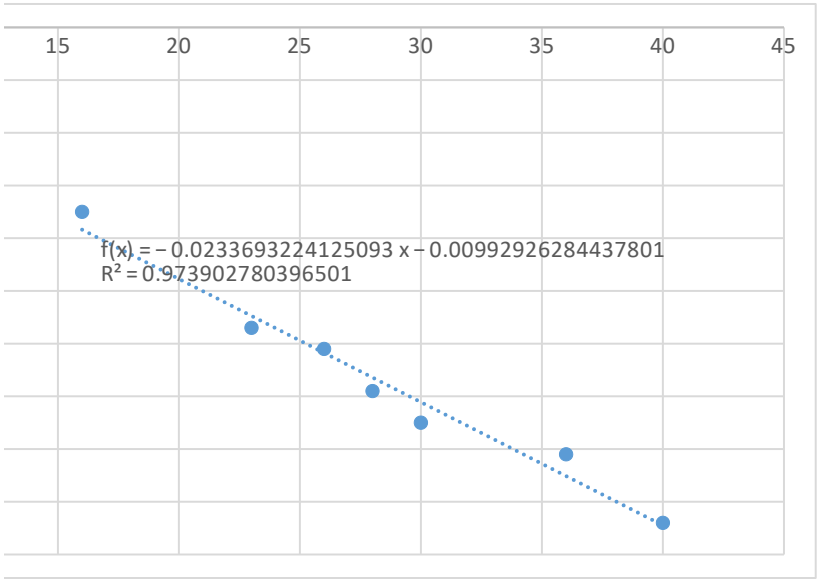

20
25
30
35
40
45
f(x) = − 0.620724841160153 ln(x) + 1.38048906518048
R² = 0.983304169001365
f(x) = − 0.0023693224125093 x − 0.00992926284437801
R² = 0.973902780396501

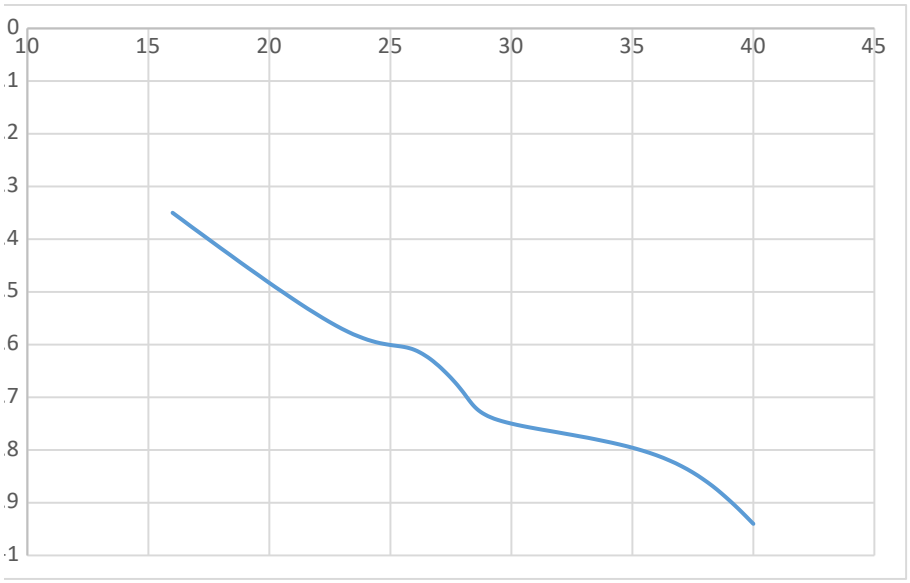

1/y = x
15
20
25
30
35
40
45
f(x) = 0.0674556200990365 x − 3.53650683831132
R² = 0.824302410002883

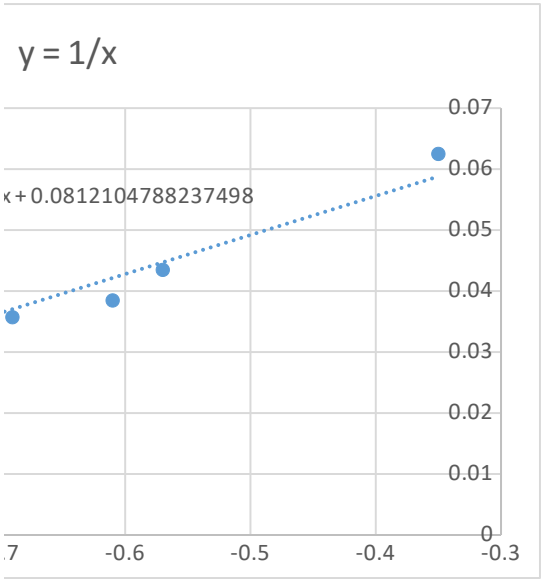

f(x) = 0.0640271516122347
R² = 0.947505534035997
-1
-0.9
-0.8


2.8
3
3.2
3.4
3.6
3.8
f(x) = − 0.620724841160153 x + 1.38048906518048
R² = 0.983304169001365




15
20
25
30
35
40
45
f(x) = − 0.0233693224125093 x − 0.00992926284437801
R² = 0.973902780396501


0
10
15
20
25
30
35
40
45
.1
.2
.3
.4
.5
.6
.7
.8
.9
-1


y = 1/x
x + 0.0812104788237498
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.7
-0.6
-0.5
-0.4
-0.3